KOLMOGOROV-SMIRNOV TEST FOR DISCRETE DISTRIBUTIONS

Mark Edward Allen

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

KOLMOGOROV-SMIRNOV TEST FOR
DISCRETE DISTRIBUTIONS

by

Mark Edward Allen

March 1976

Thesis Advisor: D. R. Barr

Approved for public release; distribution unlimited.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

Kolmogorov-Smirnov Test For Discrete Distributions

5. TYPE OF REPORT & PERIOD COVERED

Master's Thesis
March 1976

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

Mark Edward Allen

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Naval Postgraduate School
Monterey, California 93940

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Naval Postgraduate School
Monterey, California 93940

12. REPORT DATE

March 1976

13. NUMBER OF PAGES

42

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

Naval Postgraduate School
Monterey, California 93940

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

** Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Kolmogorov-Smirnov goodness-of-fit test is exact only when the hypothesized distribution is continuous, but recently Conover has extended the Kolmogorov-Smirnov test to obtain a test that is exact in the case of discrete distributions. Reasons for using this procedure instead of the regular Kolmogorov-Smirnov test when the hypothesized distribution is discrete are given. A computer subroutine is developed

to allow easy use of the procedure. The subroutine is then used to demonstrate the conservatism of the regular Kolmogorov-Smirnov test in this case and to investigate some properties of the asymptotic distributions of the test statistics.

Kolmogorov-Smirnov Test For Discrete Distributions

by

Mark Edward Allen
Lieutenant, United States Navy
B.S., University of California, Davis, 1968
M.S., University of West Florida, 1970

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 1976

ABSTRACT

The Kolmogorov-Smirnov goodness-of-fit test is exact only when the hypothesized distribution is continuous, but recently Conover has extended the Kolmogorov-Smirnov test to obtain a test that is exact in the case of discrete distributions. Reasons for using this procedure instead of the regular Kolmogorov-Smirnov test when the hypothesized distribution is discrete are given. A computer subroutine is developed to allow easy use of the procedure. The subroutine is then used to demonstrate the conservatism of the regular Kolmogorov-Smirnov test in this case and to investigate some properties of the asymptotic distributions of the test statistics.

TABLE OF CONTENTS

LIST OF FIGURES

I. INTRODUCTION

Various statistical problems reduce to the choice of a parametric form of a probability distribution of a population. A one sample goodness-of-fit test is a test of the hypothesis H_0: $F(x) = H(x)$ for all x, where F is the unknown cumulative distribution function of the population in question and H is the hypothesized cumulative distribution function. There are various test statistics that can be used in goodness-of-fit tests. The choice of which statistic to use depends on the nature of the sample, whether F is continuous or discrete, whether all of the parameters of H are known or are estimated from the sample, or whether H is a member of a certain class of distributions. The two most commonly used tests are the Chi-square and Kolmogorov-Smirnov (K-S) type goodness-of-fit tests.

The Chi-square test is based on a test statistic that is asymptotically distributed as a Chi-square random variable, and therefore is used when the sample size is relatively large. The Chi-square test does not require major assumptions on the hypothesized distribution and can be used when the parameters of the hypothesized distribution are estimated from the sample. The hypothesized distribution may be either discrete or continuous and the data may be observations of the population or grouped observations of the population.

The Kolmogorov-Smirnov test statistic has a known distribution for all sample sizes which makes the test exact. The K-S test may be preferred to the Chi-square test when the sample size is small because of the exactness of the K-S test. There is some controversy as to which of the two tests is more powerful. The relative power has been studied (see Massey, [7]) and the K-S test appears to be more powerful in some cases while the Chi-square test is more powerful in others. Traditionally, a major requirement for the K-S test has been that the hypothesized distribution, H, must be continuous. If H is not continuous, then a test of the hypothesis H_o using the traditional K-S tables is known to be conservative (see Noether, [9]).

Unfortunately, the exact degree of conservatism is not known. W. J. Conover [3] derived a method to use a K-S type test when the hypothesized distribution is discrete or when the data has already been grouped (see Darmosiswoys [5]), but the computations using this method are long and involved. In what follows, a program is developed to be used on a digital computer employing Conover's method. This program is then used to investigate the asymptotic distributions of the test statistics.

A description of notation used herein is contained in the following list:

Notation	Description
S_n	Empirical distribution function of a random sample of size n.
n	Sample size.
α	Level of significance of test.
α^*	Critical level of test.
F	Unknown distribution function of a random sample.
H	Hypothesized distribution function.
$X_1, X_2, \ldots, X_n$	Random sample of size n.
$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$	Ordered rearrangement of the random sample $X_1, \ldots, X_n$ in ascending order.
H_o	A null hypothesis in test hypotheses.
H_1	An alternate hypothesis in test hypotheses.

II. DESCRIPTION OF CONOVER'S PROCEDURE

A. KOLMOGOROV-SMIRNOV TYPE TESTS AND TEST STATISTICS

One sample K-S type tests are goodness-of-fit tests that compare the empirical cumulative distribution function of a random sample to a hypothesized cumulative distribution function. If the empirical cumulative distribution function is not close, in the sup norm sense, to the hypothesized cumulative distribution function, then the conclusion is made that the random sample did not come from the hypothesized distribution.

Let $X_1, X_2, \ldots, X_n$ be independent random variables (observations) each having the same unknown distribution F. If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represents the rearrangement of $X_1, X_2, \ldots, X_n$ in asending order, then the empirical cumulative distribution function S_n is defined by:

$$S_n(x) = \begin{cases} 0 & \text{if} \quad x < X_{(1)} \\ k/n & \text{if} \quad X_{(k)} \leq x < X_{(k+1)}, \quad k = 1,2,\ldots,n-1 \\ 1 & \text{if} \quad x \geq X_{(n)} \end{cases}$$

The K-S test may be used to test the three following hypotheses:

1. H_o: $F(x) = H(x)$ for all x

   H_1: $F(x) \neq H(x)$ for some x

2. H_o: $F(x) \geq H(x)$ for all x

   H_1: $F(x) < H(x)$ for some x

3. H_0: $F(x) \leq H(x)$ for all x

H_1: $F(x) > H(x)$ for some x

In each hypothesis, H is a specified distribution function. One of the following test statistics is used depending on the hypotheses being tested:

1. $D = \sup_x |H(x)-S_n(x)|$
2. $D^- = \sup_x (H(x)-S_n(x))$
3. $D^+ = \sup_x (S_n(x)-H(x))$

For each of the three hypotheses, a sufficiently large observation of the test statistic indicates that the null hypothesis should be rejected. If α is the level of significance desired in the test of either hypotheses 1, 2, or 3, then critical values c, c^-, or c^+ are determined as follows, according to which set of hypotheses is being tested:

1. $P(D \geq c) = \alpha$
2. $P(D^- \geq c^-) = \alpha$
3. $P(D^+ \geq c^+) = \alpha$

"P" in the above equations is the measure associated with H. If the observation d, d^-, d^+ of the statistics D, D^-, or D^+, respectively, exceeds the corresponding critical values, that null hypothesis is rejected at a level of significance of α. Instead of determining the critical values, we may compute the critical level, α^*, which is the smallest significance level at which the null hypothesis would be rejected for the

given observation d, d^-, or d^+, and compare it with α . If $\alpha^* \leq \alpha$, then the null hypothesis is rejected while if $\alpha^* > \alpha$, the null hypothesis is not rejected. The two methods are equivalent and the level of significance in both is α .

If H_o is true and H is continuous, it is known (see Darling, [4]) that the distributions of D, D^-, and D^+ are independent of H. Tables of critical values for various levels of significance of the test statistics D, D^-, and D^+ are available for use in the K-S test when H is continuous. When H is discrete, the distributions of D, D^-, and D^+ are not independent of H and the standard K-S tables cannot be used to find the critical levels of the test statistics. When H is discrete, the standard K-S tables can be used to give an approximation of the level of significance of the test because of the following demonstration. Let Y be a discrete random variable with distribution function R. If $a_1, a_2, \ldots$ are points of discontinuity of R with associated probabilities $p_1, p_2, \ldots$, then, let Z be any continuous random variable with distribution function T such that $T(a_i) - T(a_{i-1}) = p_i$, i = 1, 2,..., a_o is any point such that $a_o < a_1$. Then

$$R(a_i) = T(a_i), \quad i = 1,2,\ldots \tag{1}$$

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from R. This random sample can be thought of as having been determined by a random sample $Z_1, Z_2, \ldots, Z_n$ from T by setting $Y_k = a_i$ if $a_{i-1} < Z_k \leq a_i$, i = 1,2,..., k = 1,2,..., n. If R_n is the empirical

distribution function of $Y_1, Y_2, \ldots, Y_n$ and T_n is the empirical distribution function of $Z_1, Z_2, \ldots, Z_n$, then

$$R_n(a_i) = T_n(a_i),\ i = 1,2,\ldots \tag{2}$$

Let $D' = \sup_a \left|R_n(a)-R(a)\right|$. Since R is discrete,

$$D' = \sup_i \left|R_n(a_i)-R(a_i)\right| \ . \tag{3}$$

(1) and (2) imply $\left|R_n(a_i) - R(a_i)\right| = \left|T_n(a_i) - T(a_i)\right|$ for all $i = 1, 2, \ldots$. Then,

$$D' = \sup_i \left|R_n(a_i) - R(a_i)\right| = \sup_i \left|T_n(a_i) - T(a_i)\right| \leq$$

$$\sup_a \left|T_n(a) - T(a)\right| = D$$

which implies $P(D' \geq c) \leq P(D \geq c)$ for any c. The same argument can be used for D^- and D^+ to show that $P(D^{-'} \geq c) \leq P(D^- \geq c)$ and $P(D^{+'} \geq c) \leq P(D^+ \geq c)$. Therefore, if the standard tables are used to construct a test when H is discrete, the test is conservative.

Slakter [10] demonstrates the conservatism of the continuous K-S test when H is discrete using a computer simulation to calculate an estimate of the actual level of significance, α_k, of the hypothesis H_o where H is the discrete uniform distribution with k mass points. Ten thousand random samples were generated from the hypothesized distribution and the statistic D was evaluated. α_k was then estimated as the proportion of the ten thousand replications in which H_o was rejected. This process was repeated for various sample sizes and various k and in all cases α_k was considerably less than

the true α. For example, with k = 10, 50 observations, and $\alpha = .05$, α_k turned out to be .0166.

The use of a conservative test might at first seem desirable since it guarantees that the actual probability of rejecting the hypothesis when it is true is less than the predetermined probability of rejecting a true hypothesis. Unfortunately, this causes a decrease in the power of the test. This unknown amount of decrease in the power of the test leads us to desire that we could calculate the exact significance level of our test when H is discrete.

Since the distributions of D, D^-, and D^+ depend on H it would require a prohibitive number of tables for use in testing H_o when H is discrete, even for simple distribution families. For this reason, the use of K-S tests when H is discrete has not been investigated until recently when W. J. Conover demonstrated a method for finding the exact critical level (approximate in the two-sided case) in this instance. The program presented in this thesis makes use of Conover's procedure a practical reality.

B. CONOVER'S PROCEDURE

1. Distributions of Test Statistics

Conover derives the distribution of D, D^-, and D^+ for H continuous or discontinuous in [3]. He shows that $P(D^+ \geq t) = 1 - e_{n+1}$ where the e_i's are defined recursively as follows: $e_i = 1$ and for k = 2, 3,...,n+1

$$e_k = 1 - \sum_{j=1}^{k-1} \binom{k-1}{j-1} e_j f_j^{k-j} \qquad (4)$$

$$\text{with } f_k = P\left\{X_i < H^{-1}\left(\frac{n-k+1}{n} - t\right)\right\}, \; 1 \leq k \leq n+1 \qquad (5)$$

The X_i's are the independent identically distributed random variables with distribution function F. $H^{-1}(p)$ is defined as sup $\{x: H(x) \leq p\}$ for $0 < p \leq 1$ and as minus infinity if $p \leq 0$. If H is continuous, then with the use of the probability integral transform, it is easy to see that $f_k = 1 - \frac{k+1}{n} - t$ and (4) reduces to the form of the regular K-S statistic obtained by Birnbaum and Tingey [2]. We note that if $k > n(1-t)+1$, then from (5), $f_k = 0$ and the distribution of D^+ becomes

$$P(D^+ \geq t) = \sum_{j=1}^{m_t} \binom{n}{j-1} e_j f_j^{n-j+1} \qquad (6)$$

where m_t is the greatest integer in $n(1-t)+1$. The distribution of D^- is very similar to D^+ and is given by $P(D^- \geq t) = 1-b_{n+1}$, where the b_i's are defined recursively as follows:

$b_i = 1$ and for $k = 2,3,\ldots,n+1$

$$b_k = 1 - \sum_{j=1}^{k-1} \binom{k-1}{j-1} b_j c_j^{k-j} \qquad (7)$$

$$\text{with } c_k = P\left\{X_i > H^{-1}\left(\frac{k-1}{n} + t\right)\right\}, \; 1 \leq k \leq n+1 \qquad (8)$$

If $k > n(1-t)+1$, then $\frac{k-1}{n} + t > 1$ in (8) which implies $c_k = 0$ and the distribution of D^- becomes

$$P(D^- \geq t) = \sum_{j=1}^{m_t} \binom{n}{j-1} b_j \, c_j^{n-j+1} \tag{9}$$

$P(D \geq t)$ is approximated by $P(D \geq t) = P(D^+ \geq t) + P(D^- \geq t)$ and the following bounds for $P(D \geq t)$ are given:

$$P(D^+ \geq t) + P(D^- \geq t) - P(D^+ \geq t)\, P(D^- \geq t) \leq$$
$$P(D \geq t) \leq P(D^+ \geq t) + P(D^- \geq t) \tag{10}$$

In most tests, $P(D^+ \geq t)$ and $P(D^- \geq t)$ are small and therefore, the maximum error in this approximation is very small.

2. Calculation of Critical Levels

a. Critical Level for D^-

Let $d^- = \sup_x (H(x) - S_n(x))$ be determined from the observations. For each k such that $1 \leq k < n(1-d^-)+1$, draw a horizontal line with ordinal value of $\frac{k-1}{n} + d^-$ on the graph of H. c_k is then $1 - (\frac{k-1}{n} + d^-)$ unless the line intersects H at a discontinuity in which case c_k is one minus the height of H at the top of the jump. The b_k's are then computed from (7), and (9) is used to compute the critical level, $P(D^- \geq d^-)$.

b. Critical Level for D^+

Let $d^+ = \sup_x (S_n(x) - H(x))$ be determined from the observations. For each k such that $1 \leq k < n(1-d^+) + 1$, draw a horizontal line with ordinal value of $1 - (\frac{k-1}{n} + d^+)$

on the graph of H. f_k is then this ordinal value unless the line intersects the graph of H at a discontinuity of H in which case f_k is equal to the height of H at the bottom of the jump. The e_k's are computed using (4), and (6) is used to compute the critical level, $P(D^+ \geq d^+)$.

c. Critical Level for D

Let $d = \sup_x \left| H(x) - S_n(x) \right|$ be determined from the observations. $P(D^- \geq d)$ and $P(D^+ \geq d)$ are computed using (9) and (6) as described above, and (10) is used to put bounds on the critical level, $P(D \geq d)$.

D. SUBROUTINE "DISKS"

The calculations of critical levels as described above can be very time consuming, especially as the number of observations increases. For this reason, subroutine DISKS (Appendix A) was developed to perform these calculations. Subroutine DISKS will calculate the critical levels of equations (6) and (9) and the bounds on the critical level of D as in (10) for most discrete distributions (see Appendix A for restrictions). Subroutine DISKS was used to calculate critical levels for various examples and verified with calculations of the critical levels made by hand.

Subroutine DISKS can be modified slightly to calculate the exact size of a critical region for a test. For example, with a sample of size 10, the critical region determined from the standard tables for continuous distributions of size .1

consists of all values of D greater than .369. By inserting the value of .369 for d in a modified version of DISKS and the hypothesized distribution H, the exact size of the test when H is discontinuous (which we know is less than .1) can be calculated.

III. ASYMPTOTIC DISTRIBUTIONS OF TEST STATISTICS

A. ANALYTICAL DISTRIBUTIONS

The asymptotic distributions of D^-, D^+, and D have been studied by several people for the case when H is not continuous. Schmid [8] showed that the limiting distributions of D^-, D^+, and D do exist, but are no longer independent of H. The limiting distributions depend on the values of H at the discontinuity points. Schmid showed, for example, that if H is discontinuous at $x = x_i$, $i = 1,2,\ldots,c$, $H(x_j - 0) = f_{2j-1}$, $H(x_j) = f_{2j}$, and $f_{2c+1} = 1$, then

$$\lim_{n\to\infty} P(D < \frac{k}{\sqrt{n}}) = G(k) \text{ where}$$

$$G(k) = \sum_{i=-\infty}^{\infty} (-1)^i \left(e^{-2k^2i^2}\right) b \int \cdots \int_{A_j}$$

$$\exp\left[-\frac{1}{2} \sum_{j,m=1}^{2c} a_{jm}\, x_j x_m\right] dx_1 \ldots dx_{2c}$$

$$a_{jj} = \frac{f_{j+1} - f_{j-1}}{(f_{j+1} - f_j)(f_j - f_{j-1})}, \quad a_{j,j-1} = a_{j-1,j} = \frac{-1}{f_j - f_{j-1}}$$

$$a_{ij} = 0 \quad \text{for} \quad i < j-1 \quad \text{or} \quad i > j+1$$

$$b = (2\pi)^{-n} \prod_{j=1}^{2c+1} (f_j - f_{j-1})^{-\frac{1}{2}}$$

and

$$A_i = \bigcup_{P_1,\ldots,P_c=-\infty}^{\infty} \left\{ -k < x_{2j-1} + 2k(P_j + if_{2j-1}) < k, \quad -k < x_{2j} + 2k(P_j + kf_{2j}) < k,\ j=1,\ldots,c \right\}$$

Unfortunately, G(k) becomes undefined when H is discrete since the a's blow-up and b becomes zero. Conover [3] tried, as did this author, using the distributions of Section II to derive the asymptotic distributions, but the attempts were unsuccessful. For these reasons, a computer routine using subroutine DISKS was used to investigate the asymptotic properties of the distributions of D^-, D^+, and D. Since formulations in the literature of the limiting distributions involve multiples of the inverse of the square root of the sample size, it was decided that values of k would be determined such that $\lim_{n\to\infty} P(D \geq \frac{k}{\sqrt{n}}) = \alpha$ for various values of α. The asymptotic distributions of D^+ and D^- were not studied since they display the same basic characteristics as the asymptotic distribution of D.

B. COMPUTER PROGRAM USED

Subroutine DISKS was modified to search for the value of k such that $P(D \geq \frac{k}{\sqrt{n}})$ was as close to, but always less than, a predetermined value of α as possible. Values of n between thirty and one hundred in increments of five were used to

determine k such that $P(D \geq \frac{k}{\sqrt{n}}) = \alpha$ from (10). Values of n between eighty-five and one hundred were sometimes not used since significant errors in calculations occurred, even with double precision calculations.

The modified subroutine was used to investigate the asymptotic distribution of D when H was one of the following distributions:

1. Discrete uniform with parameter m:

$$H(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{k}{m} & k \leq x < k+1,\ k = 1,2,\ldots,\ m-1 \\ 1 & x \geq m \end{cases}$$

2. Poisson with parameter μ :

$$H(x) = \sum_{k=0}^{[x]} \frac{e^{-\mu} \mu^{k}}{k!}, \text{ where } [x] = \text{largest integer} \leq x$$

3. Geometric with parameter ρ :

$$H(x) = \sum_{k=1}^{[x]} \rho(1 - \rho)^{k-1}$$

Each distribution was investigated for various values of its respective parameter. The values of k determined for the various values of n for each particular parametric distribution were examined to determine if they appeared to be converging to some common value. The fact that the distribution of D is discrete suggested that the values of k would not converge in a uniform manner to some value, but it was hoped that, even

though it jumped around some, the convergence to a common value would be evident. By varying the values of the parameters of the various distributions, these discrete distributions would approach (in the weak convergence sense) a continuous distribution and the limiting value of k should approach the known limiting values of k for continuous distributions. For example, as m in the discrete uniform distribution increased, H has smaller and smaller jumps at each mass point and becomes "smoother" looking. If we think of the mass points being evenly distributed between zero and one, then, as the number of mass points increases, H behaves in most respects more and more like a continuous uniform distribution function between zero and one. Similarly, as the parameter of the Poisson gets larger and larger and as the parameter of the geometric gets smaller and smaller, these hypothesized cumulative distribution functions have smaller and smaller jumps at their points of discontinuity and the distribution functions get smoother and smoother. Since the usual K-S test is conservative when H is discrete, the approximating values of k for the discrete case should be always smaller than these known limiting values of k for the continuous case.

C. RESULTS

For each parametric distribution considered, as n increased, the sequence of values of k did appear to converge although,

as anticipated, not monotonically. Typical example values of k determined for various values of n are tabulated below:

n	k
30	1.095
35	1.183
40	1.107
45	1.193
50	1.131
55	1.146
60	1.162
65	1.178
70	1.165
75	1.155
80	1.148
90	1.160

These values of k were determined for the discrete uniform distribution with 10 mass points and $\alpha = .05$. The variation in k as n increases is apparent, but the value of k does appear to be fairly constant for n greater then 50. As the parameters of the three distributions were changed and the discrete distributions became "smoother" looking as described in Section III. B, the variation in k became less than that in the table above. In each parametric case that was examined, the values of k for $n > 50$ rarely varied from each other more than .03 as in the above example. The general tendency was for k to increase as n increased and then become relatively stable for $n > 50$. For $n > 50$, the smallest value k thus obtained was recorded and then all the values of k for the various values of the parameters of each distribution were plotted. Figures 1, 2, and 3 show a smooth curve approximation through the plotted k values for the three distributions with dotted lines representing the asymptotic value of k for the continuous case.

Figure 1 shows the values of k for the discrete uniform distribution for various numbers of mass points. The conservativeness of the continuous K-S test is readily apparent from this plot. For example, with twenty mass points the asymptotic k approximation is 1.16 while in the regular K-S test the asymptotic value of k is 1.36. As the number of mass points increases, the value of k is increasing toward the continuous K-S value. One of the surprising results is how slowly k converges to the continuous K-S value. Even with two hundred mass points at $\alpha = .05$, k = 1.30, which differs from 1.36 by an amount larger than expected.

Figure 2 depicts the values of k for the Poisson distribution with various values of the parameter. The curves have the same general appearance as those in Figure 1 and the same comments made about the discrete uniform apply here.

Values of k determined for the geometric distribution with various values of the parameter are plotted in Figure 3. The curves here are similar to the two preceeding distributions with the apparent convergence of the value of k to the continuous K-S value of k as the parameter decreases. With this slight modification, all of the previous comments apply here.

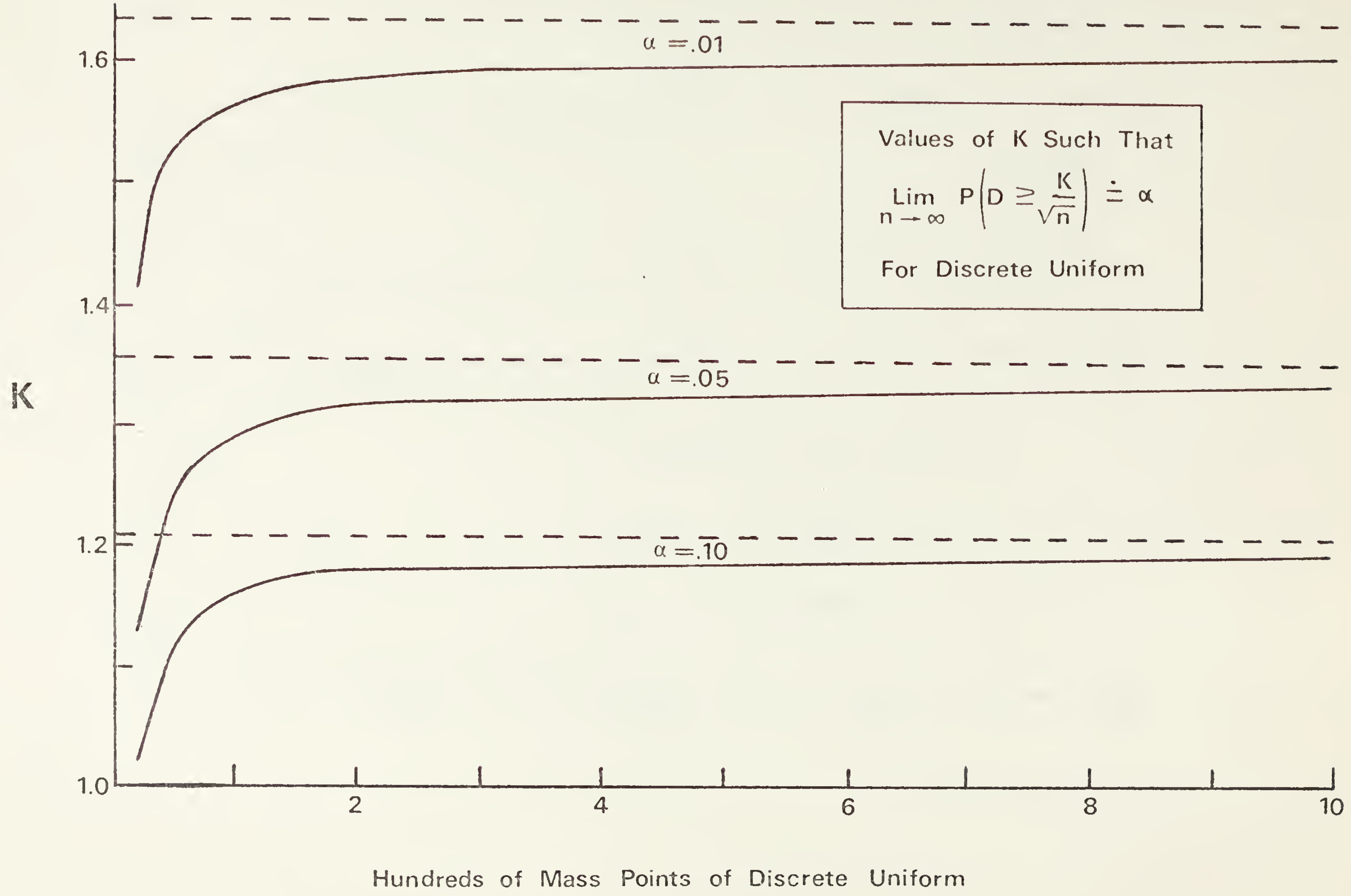
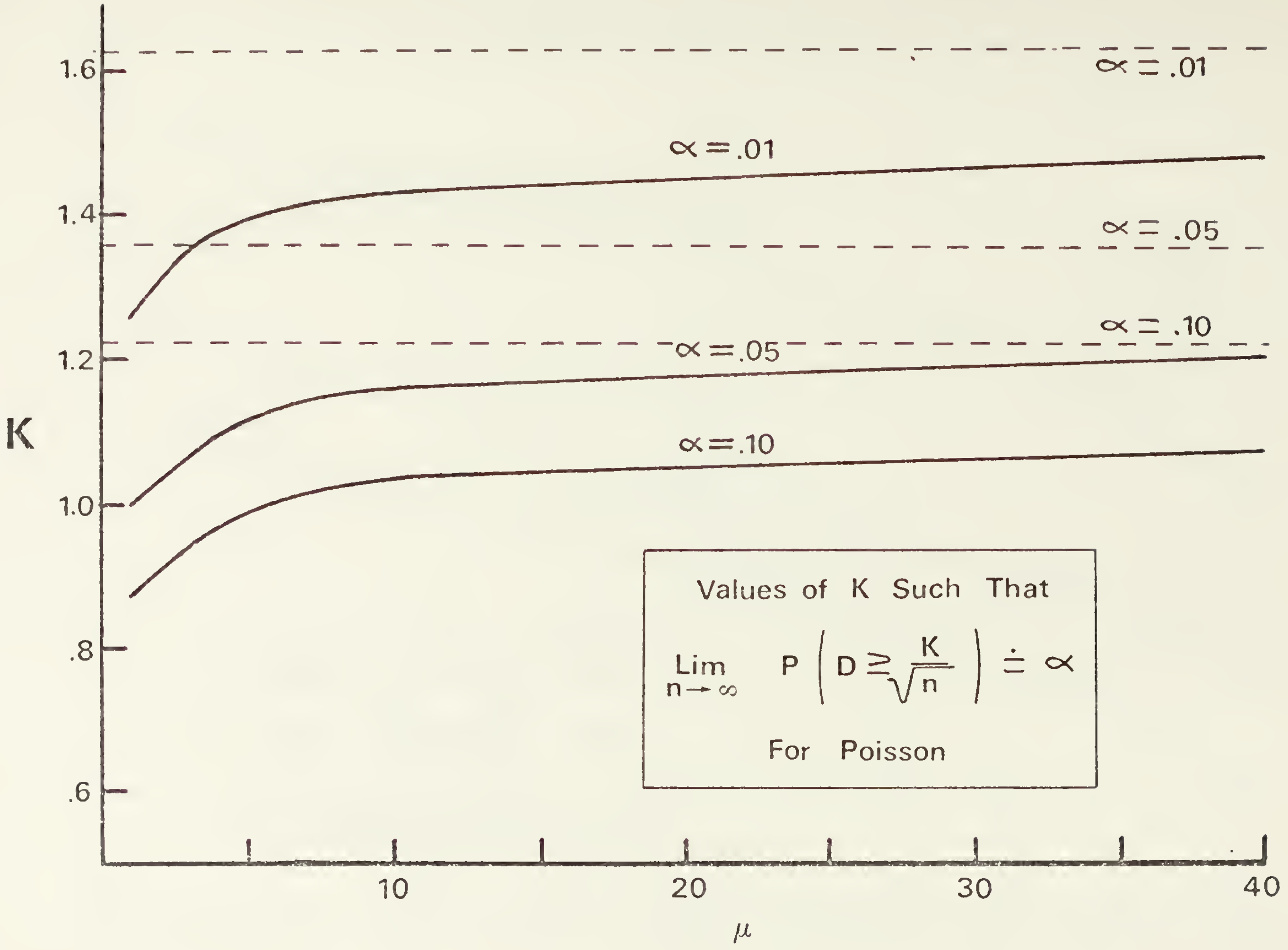

Values of K Such That
$\lim_{n \to \infty} P\left(D \geq \frac{K}{\sqrt{n}}\right) \doteq \alpha$
For Discrete Uniform
$\alpha = .01$
$\alpha = .05$
$\alpha = .10$
K
1.6
1.4
1.2
1.0
2
4
6
8
10
Hundreds of Mass Points of Discrete Uniform

FIGURE 1



1.6
1.4
1.2
1.0
.8
.6
K
$\alpha = .01$
$\alpha = .05$
$\alpha = .10$
$\alpha = .01$
$\alpha = .05$
$\alpha = .10$
Values of K Such That
$\lim_{n \to \infty} P\left(D \geq \frac{K}{\sqrt{n}}\right) \doteq \alpha$
For Poisson
10
20
30
40
μ

FIGURE 2

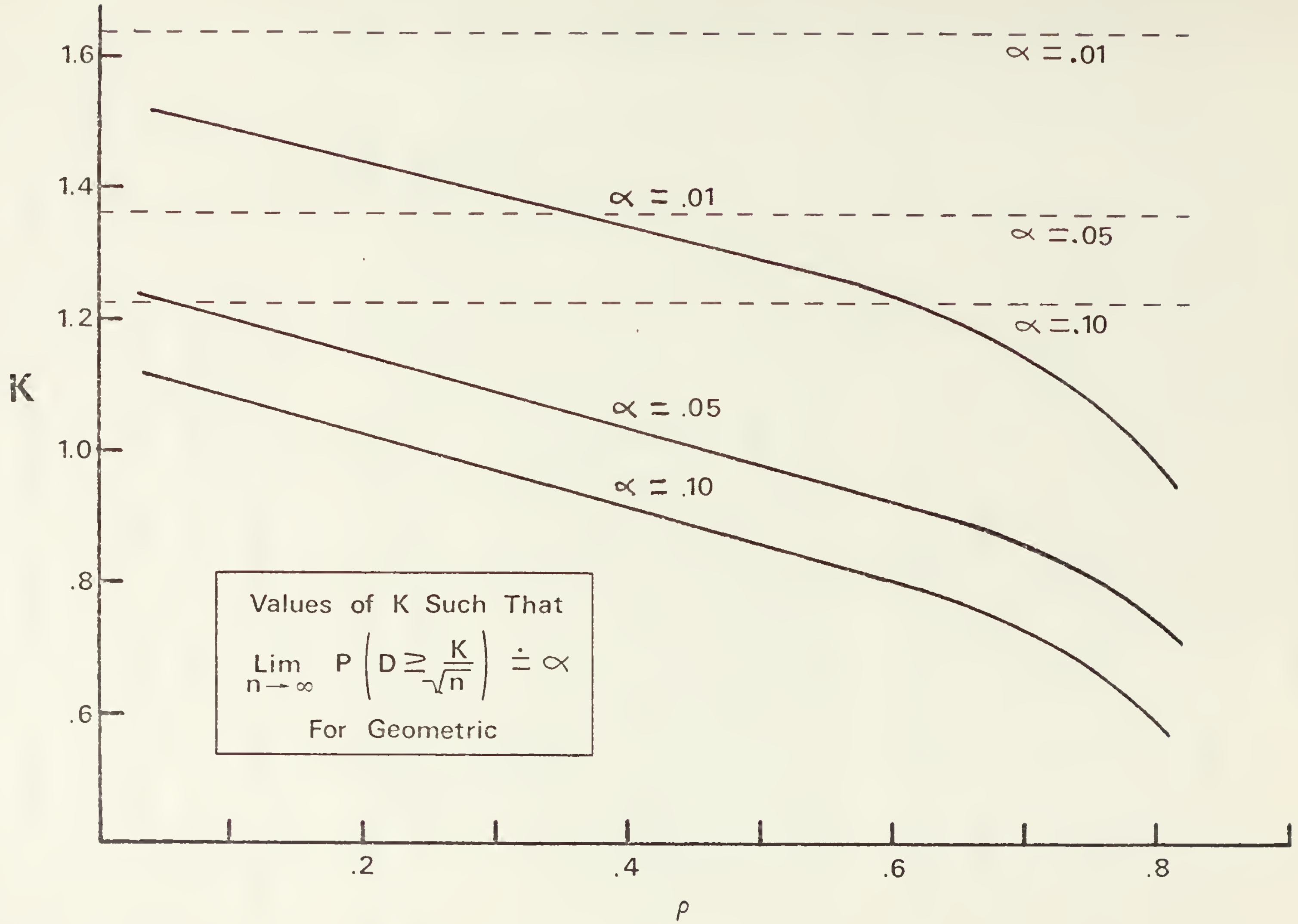

1.6
1.4
1.2
1.0
.8
.6
K
.2
.4
.6
.8
ρ
α = .01
α = .05
α = .10
α = .01
α = .05
α = .10
Values of K Such That
$\lim_{n \to \infty} P\left(D \geq \frac{K}{\sqrt{n}}\right) \doteq \alpha$
For Geometric

FIGURE 3

IV. SUMMARY AND CONCLUSIONS

1. The K-S test using the standard tabled critical values is conservative when the hypothesized distribution, H, is discrete. The test is sometimes substantially conservative as indicated in Figures 1, 2, and 3. The power of the test is reduced when the test is conservative and, therefore, it is desirable to know the exact size of a test instead of a conservative estimate.

2. Conover's procedure can be used to obtain exact (approximate in the two-sided case) critical levels for a K-S test when H is discontinuous or when the data have been grouped. The procedure can also be used to find the exact amount of conservatism of a K-S test if the standard tables are used. The only drawbacks to the procedure are the lengthy and tedious calculations required.

3. Subroutine DISKS was developed and tested to calculate the critical levels in Conover's procedure for many discrete distributions.

4. As the sample size increases, the limiting distributions of the test statistics D, D^-, and D^+ for discontinuous H exist, but, of the closed form limiting distributions investigated, they are degenerate when H is discrete. Subroutine DISKS may be modified slightly to obtain an approximation to the limiting values of k such that $P(D \geq \frac{k}{\sqrt{n}}) = \alpha$ for any $0 \leq \alpha \leq 1$.

5. The limiting values of k above were approximated as described for three distribution families. As n increased, k had a general tendency to increase and become fairly constant for $n > 50$. As the parameter of each family changed such that H had smaller jumps at mass points and become "smoother" looking, k approached the limiting value of k found in the standard K-S tables. Significantly, this convergence of k to the limiting value for the continuous case was much slower than anticipated.

6. Figures 1, 2, and 3 indicate that each family of distributions has distinctive sets of similar curves. Further investigation seems warranted to attempt to find an easy and quick means to modify the existing K-S tables for use in a K-S test when H is discrete. This would involve determining, for each family of discrete distributions, a function depending on n, α , and the parameters of the family that would modify the critical values in the standard K-S tables for continuous H into critical values for that particular family of distributions.

APPENDIX A

I. USE OF SUBROUTINE DISKS

A. PURPOSE OF SUBROUTINE

Subroutine DISKS uses Conover's [3] procedure to compute the critical level, (the probability of getting a value of the test statistic as large as the observed value when $H_0 : F(x) = H(x)$, for all x is true), of a Kolmogorov goodness-of-fit test when the hypothesized distribution is discrete. If S_n is the cumulative empirical distribution of the sample, then the following test statistics are used for the specified alternative hypothesis: (1) alternatives of the type $F = H$ use $D = \sup_x |H(x) - S(x)|$, (2) alternatives of the type $F \quad H$ use $D^- = \sup_x (H(x) - S(x))$, while (3) alternatives of the type $F \quad H$ use $D^+ = \sup_x (S(x) - H(x))$. For a given hypothesized distribution and sample of the distribution to be tested the subroutine determines the observed values of D, D^-, and D^+. If these observed values are d, d^-, and d^+, respectively, then the subroutine computes the double precision quantities PDMNS, PDPLS, PDL, and PD where:

$$\text{PDMNS} = \text{Prob}(D^- \geq d^-)$$

$$\text{PDPLS} = \text{Prob}(D^+ \geq d^+)$$

$$\text{PDL} \leq \text{Prob}(D \geq d) \leq \text{PD}$$

B. INPUT TO SUBROUTINE

1. ITYPE = 1

If all of the possible mass points of the hypothesized distribution are represented in the data, then ITYPE = 1 and the following quantities must be provided:

X -- N-dimensional vector containing the sample

H -- (M+1)-dimensional vector containing the values of the hypothesized cumulative distribution

M -- the number of distinct data points

N -- the total number of data points, less than or equal to thirty (30)

S -- a dummy vector of length (M+1)

2. ITYPE = 2

If all of the possible mass points of the hypothesized distribution are not represented, then ITYPE = 2 and the above input is modified by making X a dummy vector and S a vector of the values of the cumulative empirical distribution.

C. LIMITATIONS

The only limitation to the subroutine is that N be less than or equal to thirty (30). For N larger than thirty (30), the user need only modify the second and third dimension statements of the program by changing 30 to the number desired. The user should be cautioned that, as N gets large (about one hundred (100)), the nature of the calculations causes significant errors to propagate even with double precision calculations.

D. TIME AND CORE REQUIREMENTS

All of the times and core requirements that follow are based on runs of DISKS at W. R. Church Computer Center, Naval Postgraduate School, Monterey, California on an IBM 360/67. The subroutine requires approximately 11K of core for storage and 6.5 seconds to compile. Execution time is approximately .4 seconds for N = 10, .5 seconds for N = 20 and .55 seconds for N = 30.

E. VERIFICATION

Fifteen examples were used to verify that subroutine DISKS calculated the desired quantities correctly. In each example, the calculations were performed by hand-calculations using Conover's procedure and then compared with the computer-calculated values. Examples were formulated to exercise each "if" statement and each branching point in the subroutine at various levels of M and N. The following are three examples used in the verification process and are listed here to indicate the general types of examples used:

1. This is example 1 from Conover [3]. Let H be the discrete uniform distribution with 5 mass points on the integers 1, 2, 3, 4, 5. Suppose a random sample of size 10 with (ordered) values 1, 1, 1, 2, 2, 2, 3, 3, 3, 3 is drawn from some population. Hand-calculation shows $d^- = 0.0$, $d^+ = .4$, and $d = .4$ yielding:

$$P(D^- \geq d^-) = 1.0$$

$$P(D^+ \geq d^+) = .02081$$

$$0.04119 \leq P(D \geq d) \leq 0.04162$$

Subroutine DISKS yielded:

PDMNS = 1.0

PDPLS = 0.020809

PDL = .041184 , PD = .041617

2. This example is from Darmosiswoys [5], page 24. H has mass points 1, 2, and 3 such that P(X = 1) = .3624, P(X = 2) = .4167, and P(X = 3) = .2209 (X is a function of an exponential random variable, Y, with parameter 6.0 defined by X = 1 if $0 \leq Y \leq 2.7$, X = 2 if $2.7 < Y \leq 9.09$, and X = 3 if $Y > 9.09$). This is an example of how to handle data that has been grouped and the original sample cannot be recovered. A random sample of size 15 with values 1, 2, 3, 2, 3, 3, 1, 1, 2, 1, 3, 3, 1, 3, 3 is drawn from some population. Hand-calculation yielded:

$$.05506 \leq P(D \geq d) \leq 0.0557$$

Subroutine DISKS yielded:

PDL = 0.055174 , PD = 0.055817

3. This example illustrates how to handle discrete distributions with a countable number of mass points. Let H be the Poisson distribution with parameter 0.7. Suppose a random sample of size 10 with values 1, 3, 2, 1, 0, 1, 3, 2, 1, 2 is drawn from some population. Hand-calculations yielded:

$$P(D^- \geq d^-) = .014774$$

$$P(D^+ \geq d^+) = 0.84238$$

$$0.02316 \leq P(D \geq d) \leq 0.02386.$$

Since the number of distinct mass points is infinite, some value of M must be decided upon to use in the program. H is truncated such that all the probability associated with mass points beyond the $(M+1)^{st}$ mass point is assigned to the $(M+1)^{st}$ mass point with a corresponding grouping of sample data if necessary. With M = 4, ITYPE = 1 and $P(X>3) = 1-H(3) = .0291$ is added to $P(X = 3)$. In this case, DISKS yielded:

PDMNS = 0.014768

PDPLS = 1.0

PDL = 0.023152 , PD = 0.023277

With M = 6, ITYPE = 2 and $P(X>5) = 1-H(5) = 0.0001$ to four decimal places. In this case, DISKS yielded:

PDMNS = 0.014772

PDPLS = 0.842311

PDL = 0.023156 , PD = 0.02382

The actual hypothesized distribution is a truncated distribution, but, if the probability of all the mass points beyond the $(M+1)^{st}$ mass points is relatively small, as in the above case with M = 6, the critical levels calculated by DISKS are very good approximations to the critical levels of the untruncated hypothesized distribution.

II. SUBROUTINE TO COMPUTE CRITICAL LEVELS

```
C     ***********************************************************
C     *                                                         *
C     *SUBROUTINE DISKS(X,H,M,N,ITYPE,S,PDMNS,PDPLS,PDL,PD)*
C     *                                                         *
C     *   SUBROUTINE DISKS COMPUTES THE CRITICAL LEVELS FOR     *
C     *   THE THREE K-S STATISTICS ACCORDING TO CONOVER'S       *
C     *   PROCEDURE ( JOURNAL OF THE AMERICAN STATISTICAL       *
C     *   ASSOCIATION, SEPT.,1972, VOL 67, NO 339,PP591-6)      *
C     *   WHEN THE HYPOTHESIZED DISTRIBUTION IS DISCRETE.       *
C     *                                                         *
C     *   PARAMETERS                                            *
C     *                                                         *
C     *     X - N-DIMENSIONAL VECTOR OF DATA POINTS THAT        *
C     *         ARE REQUIRED ONLY IF ITYPE = 1                  *
C     *                                                         *
C     *     H - M+1-DIMENSIONAL VECTOR OF VALUES OF THE         *
C     *         HYPOTHEZIZED CUMULATIVE DISTRIBUTION            *
C     *         FUNCTION AT EACH DISTINCT VALUE OF X WITH       *
C     *         H(1) = 0.0 AND H(M+1) = 1.0                     *
C     *                                                         *
C     *     M - NUMBER OF DISTINCT DATA POINTS                  *
C     *                                                         *
C     *     N - NUMBER OF DATA POINTS                           *
C     *                                                         *
C     *     ITYPE - 1 IF ALL POSSIBLE MASS POINTS ARE           *
C     *               REPRESENTED IN THE DATA                   *
C     *                                                         *
C     *             2 IF NOT ALL POSSIBLE MASS POINTS ARE       *
C     *               REPRESENTED                               *
C     *                                                         *
C     *     S - VALUES OF THE EMPIRICAL DISTRIBUTION            *
C     *         FUNCTION AT MASS POINTS.  INPUT ONLY IF         *
C     *         ITYPE = 2                                       *
C     *                                                         *
C     *     PDMNS - DOUBLE PRECISION OUTPUT CRITICAL LEVEL      *
C     *             FOR D-MINUS                                 *
C     *                                                         *
C     *     PDPLS - DOUBLE PRECISION OUTPUT CRITICAL LEVEL      *
C     *             FOR D-PLUS                                  *
C     *                                                         *
C     *     PDL - DOUBLE PRECISION OUTPUT LOWER BOUND ON        *
C     *           CRITICAL LEVEL FOR D                          *
C     *                                                         *
C     *     PD - DOUBLE PRECISION OUTPUT UPPER BOUND ON         *
C     *          CRITICAL LEVEL FOR D                           *
C     *                                                         *
C     *   USAGE - REJECT HYPOTHESIS F(X) = H(X) IF PREDE-       *
C     *           TERMINED CRITICAL LEVEL IS GREATER THAN PD*
C     *                                                         *
C     *           REJECT HYPOTHESIS F(X) GREATER THAN H(X)      *
C     *           IF PREDETERMINED CRITICAL LEVEL IS            *
C     *           GREATER THAN PDMNS                            *
C     *                                                         *
C     *           REJECT HYPOTHESIS F(X) LESS THAN H(X) IF      *
C     *           PREDETERMINED CRITICAL LEVEL IS GREATER       *
C     *           THAN PDPLS                                    *
C     *                                                         *
C     ***********************************************************
C
      SUBROUTINE DISKS  (X,H,M,N,ITYPE,S,PDMNS,PDPLS,PDL,PD)
      DIMENSION X(N),H(N),S(N),CO(30,30),J(30),F(30),CD(30)
      DIMENSION B(30), E(30), BD(30), ED(30), C(30), FD(30)
      REAL*8 CO,F,CD,FD,B,E,BD,ED,C,BSUM,ESUM,PDMNS,PDPLS
      REAL*8 PDM,PDP,Y,PDL,PD
      NM1 = N-1
```

```
      RN = FLCAT(N)
      DMNS = 0.0
      DPLS = 0.0
      MP1 = M+1
      EFS = 1.E-6
      IF (ITYPE.EQ.2) GO TO 8
C
C     ******************************************************
C     *                                                    *
C     *  SORT X'S IN ASCENDING ORDER. J IS SORTED INDEX    *
C     *                                                    *
C     ******************************************************
C
      DC 1 K1=1,N
      J(K1) = K1
    1 CCNTINUE
C
C
      DO 3 K2=1,NM1
      IY = K2+1
C
      DC 2 K3=IY,N
      IF (X(J(K2)).LE.X(J(K3))) GO TO 2
      IDUM = J(K2)
      J(K2) = J(K3)
      J(K3) = IDUM
    2 CCNTINUE
C
    3 CCNTINUE
C
C     ******************************************************
C     *                                                    *
C     *  COMPUTE EMPIRICAL DISTRIBUTION FUNCTION, S        *
C     *                                                    *
C     ******************************************************
C
      S(1) = 0.0
      SUM = 0.0
      K = 2
      I = 1
    4 IY = I+1
C
      DC 5 K4=IY,N
      IF (X(J(K4)).GT.X(J(I))) GO TO 6
    5 CCNTINUE
C
    6 I = K4
      SUM = SUM+(K4-IY+1)/RN
      S(K) = SUM
      K = K+1
      IF (K4.EQ.N) GO TO 7
      GC TO 4
    7 S(K) = 1.0
C
C     ******************************************************
C     *                                                    *
C     *  COMPUTE DPLS, DMNS, AND D                         *
C     *                                                    *
C     ******************************************************
C
    8 DC 9 K17=2,M
      DIFF = H(K17)-S(K17)
      DIFF2 = -DIFF
      IF (DMNS.LT.DIFF) DMNS=DIFF
      IF (DPLS.LT.DIFF2) DPLS=DIFF2
    9 CCNTINUE
C
      D = DMNS
      IF (DPLS.GT.D) D = DPLS
      NMNS = RN*(1.0-DMNS)+0.9999
      NFLS = RN*(1.0-DPLS)+0.9999
      ND = RN*(1.0-D)+0.9999
```

```
C     *****************************************************
C     *                                                   *
C     *  COMPUTE C'S AND F'S                              *
C     *                                                   *
C     *****************************************************
C
      NC = 1
C
      DO 14 K18=1,NMNS
      ORD = DMNS+(K18-1.0)/RN
C
      DO 10 K19=NC,MP1
      IF (ORD.LT.H(K19)) GO TO 11
   10 CONTINUE
C
   11 IY = K19-1
      OMH = ORD-H(IY)
      IF (ABS(OMH).LE.EPS) GO TO 12
      C(K18) = 1.0-H(K19)
      GO TO 13
   12 C(K18) = 1.0-ORD
   13 NC = IY
   14 CONTINUE
C
      NC = 1
C
      DO 19 K20=1,NPLS
      ORD = 1.0-DPLS-(K20-1.0)/RN
C
      DO 15 K21=NC,MP1
      NB = MP1-K21+1
      IF (ORD.GT.H(NB)) GO TO 16
   15 CONTINUE
C
   16 IY = NB+1
      HMO = H(IY)-ORD
      IF (ABS(HMO).LE.EPS) GO TO 17
      F(K20) = H(NB)
      GO TO 18
   17 F(K20) = ORD
   18 NC = MP1-NB
   19 CONTINUE
C
C     *****************************************************
C     *                                                   *
C     *  COMPUTE CD'S AND FD'S                            *
C     *                                                   *
C     *****************************************************
C
      NC = 1
C
      DO 24 K22=1,ND
      ORD = D+(K22-1.0)/RN
C
      DO 20 K23=NC,MP1
      IF (ORD.LT.H(K23)) GO TO 21
   20 CONTINUE
C
   21 IY = K23-1
      OMH = ORD-H(IY)
      IF (ABS(OMH).LE.EPS) GO TO 22
      CD(K22) = 1.0-H(K23)
      GO TO 23
   22 CD(K22) = 1.0-ORD
   23 NC = IY
   24 CONTINUE
C
      NC = 1
C
      DO 29 K24=1,ND
      ORD = 1.0-D-(K24-1.0)/RN
C
```

```
      DO 25 K25=NC,MP1
      NB = MP1-K25+1
      IF (ORD.GT.H(NB)) GO TO 26
   25 CONTINUE
C
   26 IY = NB+1
      HMO = H(IY)-ORD
      IF (ABS(HMO).LE.EPS) GO TO 27
      FD(K24) = H(NB)
      GO TO 28
   27 FD(K24) = ORD
   28 NC = MP1-NB
   29 CONTINUE
C
C     ********************************************************
C     *                                                      *
C     *  COMPUTE CO(I,J), COMBS I-1 TAKEN J-1 AT A TIME      *
C     *                                                      *
C     ********************************************************
C
      NP1 = N+1
C
      DO 31 I=2,NP1
      CO(I,1) = 1.0
      IM1 = I-1
C
      DO 30 JJ=2,I
      JM1 = JJ-1
      CO(I,JJ) = (CO(I,JM1)*(I-JJ+1.0))/(JJ-1)
   30 CONTINUE
C
   31 CONTINUE
C
C     ********************************************************
C     *                                                      *
C     *  COMPUTE B'S, E'S, BD'S, AND ED'S                    *
C     *                                                      *
C     ********************************************************
C
      B(1) = 1.0
C
      DO 33 K26=2,NMNS
      BSUM = 1.0
      IY = K26-1
C
      DO 32 K27=1,IY
      BSUM = BSUM-CO(K26,K27)*(C(K27)**(K26-K27))*B(K27)
   32 CONTINUE
C
      B(K26) = BSUM
   33 CONTINUE
C
      E(1) = 1.0
C
      DO 35 K28=2,NPLS
      ESUM = 1.0
      IY = K28-1
C
      DO 34 K29=1,IY
      ESUM = ESUM-CO(K28,K29)*(F(K29)**(K28-K29))*E(K29)
   34 CONTINUE
C
      E(K28) = ESUM
   35 CONTINUE
C
      BD(1) = 1.0
      ED(1) = 1.0
C
      DO 37 K30=2,ND
      BSUM = 1.0
      ESUM = 1.0
      IY = K30-1
```

```
      DO 36 K31=1,IY
      BSUM = BSUM-CO(K30,K31)*(CD(K31)**(K30-K31))*BD(K31)
      ESUM = ESUM-CO(K30,K31)*(FD(K31)**(K30-K31))*ED(K31)
   36 CONTINUE
C
      BD(K30) = BSUM
      ED(K30) = ESUM
   37 CONTINUE
C
C     ***************************************************************
C     *                                                             *
C     *   COMPUTE CRITICAL LEVELS, PDMNS, PDPLS, AND PD             *
C     *                                                             *
C     ***************************************************************
C
      PDMNS = 0.0
      PDPLS = 0.0
      PDP = 0.0
      PDM = 0.0
C
      DO 38 K32=1,NMNS
      PDMNS = PDMNS+CO(NP1,K32)*B(K32)*(C(K32)**(N-K32+1))
   38 CONTINUE
C
C
      DO 39 K33=1,NPLS
      PDPLS = PDPLS+CO(NP1,K33)*E(K33)*(F(K33)**(N-K33+1))
   39 CONTINUE
C
C
      DO 40 K34=1,ND
      IY = N-K34+1
      Y = CO(NP1,K34)
      PDM = PDM+Y*BD(K34)*(CD(K34)**IY)
      PDP = PDP+Y*ED(K34)*(FD(K34)**IY)
   40 CONTINUE
C
      PD = PDM+PDP
      PDL = PD-PDM*PDP
      RETURN
      END
```

LIST OF REFERENCES

1. Anderson, T. W. and Darling, D. A., "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes," Annals of Mathematical Statistics, v. 23, p. 193-212, 1952.

2. Birnbaum, Z. W. and Tingey, F. H., "One-Sided Confidence Contours for Probability Distribution Functions," The Annals of Mathematical Statistics, v. 22, p. 592-596, December, 1951.

3. Conover, W. J., "A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions," Journal of the American Statistical Association, v. 67, p. 591-596, September, 1972.

4. Darling, D. A., "The Kolmogorov-Smirnov, Cramer-Von Mises Tests," Annals of Mathematical Statistics, v. 28, p. 823-838, 1957.

5. Darmosiswoys, S., Kolmogorov-Smirnov Test for Grouped Data, M. S. Thesis, Naval Postgraduate School, Monterey, California, 1975.

6. Doob, J. L., "Heuristic Approach to the Kolmogorov-Smirnov Theorems," Annals of Mathematical Statistics, v. 20, p. 393-403, 1949.

7. Massey, F. J., "The Kolmogorov-Smirnov Test for Goodness-of-Fit," Journal of the American Statistical Association, v. 46, p. 68-78, 1951.

8. Noether, G. E., "Note on the Kolmogorov Statistic in the Discrete Case," Metrika, v. 7, No. 2, p. 115-116, 1963.

9. Schmid, P., "On the Kolmogorov and Smirnov Limit Theorems for Discontinuous Distribution Functions," The Annals of Mathematical Statistics, v. 29, p. 1011-1027, December, 1958.

10. Slakter, M. J., "A Comparison of the Pearson Chi-squared and Kolmogorov Goodness-of-Fit Tests with Respect to Validity," Journal of the American Statistical Association, v. 60, p. 854-858, September, 1965.

11. Walsh, J. E., "Bounded Probability Properties of Kolmogorov-Smirnov and Similar Statistics for Discrete Data," Annals of the Institute of Statistical Mathematics, v. 15, No. 2, p. 153-158, 1963.

INITIAL DISTRIBUTION LIST

No. Copies

1. Defense Documentation Center 2
Cameron Station
Alexandria, Virginia 22314

2. Library, Code 0212 2
Naval Postgraduate School
Monterey, California 93940

3. Department Chairman, Code 55 2
Department of Operations Research and
Administrative Science
Naval Postgraduate School
Monterey, California 93940

4. Associate Professor D. R. Barr, Code 55 Bn 1
Department of Operations Research and
Administrative Science
Naval Postgraduate School
Monterey, California 93940

5. Associate Professor F. R. Richards, Code 55 Ri 1
Department of Operations Research and
Administrative Science
Naval Postgraduate School
Monterey, California 93940

6. Lt. Mark Edward Allen, USN 1
c/o Mrs. Lois Rollins
15 Lincoln Avenue
Woodland, California 95695